

Research Article

Yanjiao Zhu and Jiehui Hu*

The effects of watching subtitled videos on the perception of L2 connected speech by L1 Chinese-L2 English speakers

<https://doi.org/10.1515/phon-2023-0011>

Received May 17, 2023; accepted January 21, 2024; published online February 19, 2024

Abstract: The current study explores whether watching subtitled videos could facilitate L1 Chinese-L2 English speakers' perception of L2 English connected speech. Three hundred ninety seven Chinese college students of L2 English completed a video-based spot dictation task after watching English videos with or without L1/L2 subtitles, featuring various connected speech types (e.g., linking, deletion, and their combinations). Results suggested an overall facilitation effect of watching videos on L2 connected speech perception, which was modulated by proficiency, subtitle form, and the complexity of connected speech. First, subtitled videos were more facilitative than non-subtitled videos in L2 perception. Second, participants with higher L2 proficiency better perceived English connected speech than those with lower proficiency. Third, the more connective devices an item used, the more difficult it was for L2 perception. When this complexity was controlled, the L2 perception was not influenced by connected speech type. Finally, the complexity of connected speech also mediated the subtitle facilitation effects. When the connected speech involved triple connective devices, L2 speakers benefited more from L1 subtitles than L2 subtitles. The findings can provide insights into multi-modal speech perception and English connected speech learning.

Keywords: connected speech perception; L2 acquisition; video comprehension

*Corresponding author: Jiehui Hu, University of Electronic Science and Technology of China, Chengdu, P.R. China, E-mail: jerrywho@163.com

Yanjiao Zhu, University of Electronic Science and Technology of China, Chengdu, P.R. China, E-mail: zhuyan.jiao@hotmail.com. <https://orcid.org/0000-0003-3865-6067>

1 Introduction

Connected speech is a continuous sequence of utterances occurring in natural conversations. English is a language with rich types of connected speech devices (Alameen and Levis 2015). For instance, the change from “*could have*” to “*could’ ave*” involves the deletion of /h/, while the change from “*do not*” to “*don’t*” is due to the contraction of two words (see below for a further description of connected speech types). English connected speech is difficult for Chinese speakers to master due to the prosodic difference between the two languages. Mandarin Chinese speech is articulated in a staccato way, with noticeable syllable boundaries (Duanmu 2007), while English speech is more like singing in a legato manner, with sounds smoothly connected to each other (Roach 2009). Since Mandarin Chinese uses fewer connected speech devices than English, Chinese speakers of L2 English are identified to perform poorly in both pronouncing and understanding English connected speech (e.g., Liang 2015; Mao and Chen 2013). Meanwhile, connected speech plays a pivotal role in L2 speech perception, as it usually occurs on high-frequency function words that L2 listeners heavily rely on (e.g., Field 2008; Walker 2010). Therefore, efforts should be taken to improve L2 speakers’ connected speech skills.

Mounting evidence has suggested the potential beneficial effects of subtitled videos on L2 speakers’ vocabulary learning (Bird and Williams 2002; Koolstra and Beentjes 1999; Teng 2022) and overall listening comprehension (Rodgers and Webb 2017; Winke et al. 2010), but it remains uncertain whether the aforementioned benefits can be extrapolated to connected speech perception. Only a few studies have focused on the effects of watching subtitled videos on L2 connected speech perception, and the findings are promising. Yang and Chang (2014) found that watching L2 subtitled videos could facilitate the perception of English connected speech by a group of Taiwanese speakers of L2 English with homogeneous L2 proficiency. Chen et al. (2021) suggested that video-based instruction improved a class of Taiwanese L2 English speakers’ learning attitudes and accuracies in connected speech production, but the study did not mention whether the videos used L1 or L2 subtitles. It is not clear whether this facilitation effect can be extended to videos of any subtitle form and L2 speakers of different proficiencies. In fact, previous studies have shown that the effectiveness of subtitled videos on L2 learning is dependent on language proficiency and subtitle form (Leveridge and Yang 2013; Lwo and Lin 2012). In addition, L1 Chinese speakers of L2 English were observed to exploit some types of connected speech (e.g., assimilation) better than others (e.g., reduction) (Liang 2015; Mao and Chen 2013), so the type of connected speech may also need further exploration. The current study examines whether watching subtitled videos can facilitate the perception of L2 connected speech by Chinese speakers of L2 English and

whether the effect is modulated by connected speech type, subtitle form, and language proficiency. Typical perceptual errors were also analyzed to reveal the processing mechanism of L2 connected speech in videos.

1.1 The acquisition of L2 English connected speech by Chinese L1 speakers

English connected speech can involve a wide range of phonetic processes that make words to sound differently from their citation forms. According to Alameen and Levis (2015), connected speech can be divided into six types, including linking, deletion, insertion, modification, reduction, and contraction. Linking makes two words sound like one single word through resyllabification, such as the changing of “*some of*” [sʌm. əv] into [sʌ.məv]. Deletion is the loss of segments, like the loss of /h/ in “*Did he do his homework?*”. Insertion is adding extra sounds to link two vowels across two words, such as the addition of the glide /r/ in “*the idea(r) of*”. Modification is changing the pronunciation of a segment, which may include palatalization (e.g., “*did you*” pronounced as [dɪdʒu] rather than [dɪdju]), assimilation (e.g., “*sun beam*”, where /n/ is pronounced as [m] before a bilabial stop), flapping (e.g., “*went out*”, where /t/ is pronounced as an alveolar flap) and glottalization (e.g., “*that car*” is pronounced as [ðætʔkɑɪ] rather than [ðætkaɪ]). Reduction is the lenition of vowels and consonants in unstressed positions such as the reduction of “*to you*” [tu ju] into [tə jə]. Contraction includes lexical chunks like “*gonna*” and contractions like “*they’re*”. Due to such rich phonetic modifications and combinations of them, English words produced in connected speech can differ considerably from the ones in citation speech, especially when spoken in fast speech and colloquial style.

English connected speech is difficult to learn for Chinese speakers for at least two reasons. First, no matter how it is implemented by native speakers, most language teachers use carefully and clearly spoken speech to ensure the comprehensibility of their language, which could result in learners’ inability to understand the connected speech in colloquial English (Chen et al. 2021). Second, connected speech is represented differently in Chinese and English. Mandarin Chinese is a syllable-timed language with all the syllables uttered with approximately equal stress and duration (Duanmu 2007), which means that Chinese connected speech seldom uses syllable reduction. In comparison, English is a stress-timed language with each stressed syllable occurring at regular intervals (Roach 2009), so English connected speech employs various devices such as linking, reduction, and syllabification to achieve this stress pattern. Thus, English connected speech contains many phonetic modification processes that are unfamiliar to Chinese speakers of L2 English.

In various studies, Chinese speakers of L2 English showed substandard performances in English connected speech perception and production. Liang (2015) found that Chinese university sophomores majoring in English performed poorly in comprehending connected speech. Among all the tested connected speech types, assimilation and linking were better perceived than reduction. The difficulty in acquiring English reduction was supported by Mao and Chen (2013), demonstrating that Chinese speakers of L2 English did not produce the English reduced schwa /ə/ in their L2 speech. Similar observations were made for L2 English speakers in Hong Kong with L1 Cantonese, a Chinese variety that also seldom uses connected speech devices. Setter et al. (2014) examined the perception of English “junctures” (connected speech across grammatical boundaries) in ambiguous word pairs (e.g. “*great eyes*” vs. “*grey ties*”). The study observed that the Hong Kong listeners were only able to correctly identify 59.58 % of junctures in British English. Wong et al. (2021a) studied the L2 production of various English connected speech and suggested that L2 speakers used too much deletion and not enough linking and reduction. In another study, Wong et al. (2021b) used an error-based approach to analyze the “slips of the ear” in L2 perception of English connected speech. Results from a dictation task suggested that Cantonese L1 speakers tended to replace vowels and consonants, change the number of syllables, and substitute function words.

The above studies collectively suggest that Chinese speakers had difficulty producing and perceiving L2 English connected speech. This difficulty may be modulated by the type of connected speech, but there has not been any study that evaluates the influence of connected speech type on L2 acquisition difficulty. Besides, findings of the L2 perceptual errors were limited to Cantonese speakers, and Mandarin speakers’ patterns await further exploration.

1.2 The processing of subtitled videos

Subtitled videos provide textual, audio, and visual information about the speech. At first glance, the multi-modal input may overload viewers, but various information processing theories suggest otherwise. The Dual Coding Theory (Paivio 1986, 2007) believes that verbal and visual systems are functionally independent but interconnected, and the two systems work in support of each other so as to facilitate information recall. The Cognitive Theory of Multimedia Learning (Mayer 2005) posits that the simultaneous presentation of words and pictures can help L2 speakers make connections and achieve better learning outcomes. The Cognitive Load Theory (Sweller et al. 2011) proposes that the potential cognitive load results from the inability to process information all at once through a verbal or a visual channel, as the working memory is limited. Subtitled videos present information in audio,

visual, and textual channels which are complementary to each other, and the distribution of information in three channels could balance the processing load in each channel so as to reduce the cognitive load (Vanderplank 2016).

However, the facilitative effect of subtitled English videos on L2 speech perception is not always constant. The first possible contributing factor is the subtitle form. Hayati and Mohmedi (2011) examined the comprehension of short DVD videos by 90 intermediate L1 Persian – L2 English speakers, grouped in L2 English subtitle, L1 Persian subtitle, and no subtitle conditions. Their results suggested that the L2 English subtitle group performed considerably better than the L1 Persian subtitle group, which in turn performed better than the no subtitle group in listening comprehension tests, so the study showed the superiority of L2 subtitles over L1 subtitles. On the contrary, Markham et al. (2001) found the advantage of L1 subtitles over L2 subtitles. The study investigated the comprehension of Spanish DVD videos by 169 intermediate-level L2 Spanish speakers with L1 English, under L1 English, L2 Spanish, and no subtitle conditions. Results showed that the L1 English subtitle group outperformed the L2 Spanish subtitle group and the no subtitle group.

L2 proficiency is another potential variable. Pujolà (2002) observed that lower-level students perceived L2 subtitles as a more necessary aid than higher-level students in video comprehension. The same pattern was evidenced in Leveridge and Yang (2013), which employed a Caption Reliance Test and showed a negative correlation between subtitle reliance and L2 proficiency. In contrast, Montero Perez et al. (2014) found that higher-level L2 speakers benefited more from L2 subtitles, evidenced by a positive correlation between their vocabulary size and captioned video comprehension. Other studies did not find proficiency effects. In Montero Perez et al. (2013), proficiency level did not affect the effect sizes of L2 subtitles on listening comprehension and vocabulary learning. Similar findings were reported in Winke et al. (2010) where the effectiveness of subtitles stayed constant across all proficiency levels. Given the inconsistent findings, Gass et al. (2019) alluded to a proficiency zone within which L1 subtitles were most useful, a proficiency range where the L2 speakers neither find the videos too far beyond or below their comprehension.

The above studies show that although watching subtitled videos can benefit L2 speech perception, the effects may differ with L2 proficiency and subtitle form. Due to the inconsistent findings in previous studies, attention should be paid to L2 proficiency and subtitle form factors.

1.3 The effect of watching subtitled videos on L2 connected speech production and perception

Adopting various designs, a few studies showed that watching subtitled videos could overall improve Chinese speakers' L2 English connected speech skills. Yang and Chang (2014) tested 44 L1 Chinese-L2 English university students' comprehension of English connected speech in animation, cartoons, movies, and TV series under full, keyword-only, and annotated keyword L2 subtitle conditions using a pretest-posttest paradigm. The full subtitle condition displayed the complete transcription of the speech, the keyword condition presented the meaningful words only, and the annotated keyword condition showed keywords along with symbols (e.g., a blue dot to mark an assimilated sound, and a grey letter to mark a reduced sound). The L2 speakers were found to improve their listening comprehension scores after watching these subtitled videos, and the greatest improvement was observed for the annotated keyword L2 subtitle condition. As the study focuses on subtitle form, it is unknown which connected speech type is most difficult to learn.

Chen et al. (2021) explored the utility of watching subtitled videos in L2 English connected speech instruction. Forty-eight Taiwanese English L2 speakers participated in a 7-week video-based treatment in classrooms. Each week, the teacher showed students short YouTube videos presenting the main features of connected speech and explained the relevant connected speech rules. Before and after the treatment, the students completed a questionnaire that examined their attitudes toward English learning, as well as a reading-aloud task that examined their connected speech accuracy. Results found that students demonstrated enhanced learning attitudes towards English speaking. The students also improved their speaking skills after the instruction. The improvement was greatest in the linking of /k/ and /a/ at 31 %, while other types of linking such /f/ and /o/ showed moderate improvement at 10–30 %. It should be noted that this study only trained English linking skills, so the conclusion may not be extended to other types of connected speech.

Moreover, Wong et al. (2020) specifically revealed the influences of subtitle form and word frequency on L2 English connected speech perception. Focusing on three types of connected speech, namely assimilation, deletion, and linking, the study tested 28 Hong Kong teenage L2 English speakers' decoding of English minimal pairs of connected speech (e.g., *“ten coins”* vs. *“tank coins”*). By manipulating two parameters, namely the link between the content of the spoken word phrases and the subtitles (matched vs. mismatched) and the degree of word frequency (high-frequency vs. low-frequency), the study found that both matched- and

mismatched subtitles could facilitate the perception of L2 connected speech, and that L2 speakers performed better on connected speech items with high word frequency.

In summary, Yang and Chang (2014) and Chen et al. (2021) showed that watching L2 subtitled videos could help L2 speakers perceive and produce English connected speech. Wong et al. (2020) reached a similar conclusion and added the effects of word frequency. Yet there are still several remaining areas worth exploring. First, using controlled stimuli, the above studies investigated assimilation, elision, and linking separately, whereas in real life, an utterance can contain several types of connected speech at the same time. It is unknown whether similar observations could be made for such complex utterances. Second, previous audio-only studies (e.g., Liang 2015) found that reduction was more challenging than assimilation and linking for L2 English speakers with L1 Mandarin. They even moved one step forward and pinpointed the common perceptual errors (Wong et al. 2021a, 2021b). However, it is unclear whether these findings could be extended to video-based multi-model speech perception. What is more, the majority of past studies examined L2 subtitles, albeit in various forms, but it is common practice for L2 speakers to watch L1-subtitled English videos, and the remaining question is whether L1 subtitles and L2 subtitles are equally useful in helping L2 speakers perceive connected speech in videos. Finally, none of the above connected speech studies have addressed the language proficiency factor, which was widely mentioned in previous video processing studies (e.g., Leveridge and Yang 2013). Hence, the study proposes the following research questions:

Research Questions (RQs):

1. Does watching subtitled English videos facilitate L1 Chinese-L2 English speakers' perception of English connected speech? If so, are L2 subtitled videos more beneficial than L1 subtitled videos?
2. Do language proficiency and the type of connected speech affect L1 Chinese-L2 English speakers' perception of English connected speech in videos? If so, what are their effects?
3. What kind of errors do L2 speakers commit during the perception of English connected speech in videos? What could be the possible causes for these errors?

To answer RQ 1, we tested L2 speakers' perception of English connected speech in videos under L1 Chinese subtitle, L2 English subtitle, and no subtitle conditions using a spot dictation task. A control group of L1 Chinese-L2 English speakers took the same perception test without watching any videos. A comparison between groups and across conditions would elucidate the effects of watching different forms of subtitled videos on connected speech perception. To examine the proficiency effects in RQ 2, we assessed the relationship between L2 speakers' connected speech performance and their scores in a widely recognized national English proficiency test.

To find out the effects of connected speech type, we selected a range of naturally spoken connected speech phrases, which enabled us to compare the performances of different types of connected speech and pinpoint the acquisition difficulty that L2 speakers faced. To answer RQ3, we further examined L2 speakers’ written answers in the spot dictation task and discussed their common perceptual errors.

2 Methods

2.1 Participants

Three hundred ninety seven L1 Chinese-L2 English bilingual students at the University of Electronic Science and Technology of China participated in the study. They were randomly assigned tasks to form an experimental group of 320 students and a control group of 77 students. The experimental group was further randomly divided into four categories according to the version of videos they watched (see Section 2.2 for video versions). Table 1 lists out the age, gender, age of L2 acquisition, and L2 proficiency of each participant group.

As shown in Table 1, the groups of L2 speakers had similar demographic and language backgrounds. They were recruited from General English courses, offered to first-year university students who specialized in STEM (science, technology, engineering, and mathematics). Since over 90 % of the students at this university were males, the participants of this study were gender-imbalanced. The uneven gender distribution was not expected to seriously invalidate our findings, because many studies showed that men and women could perceive L2 speech sounds equally well (Bryła-Cruz 2021; Ceuleers et al. 2021), despite the gender difference in L2 acquisition due to social or cultural factors (Moyer 2016; Pavlenko and Piller 2008). The participants were comparable in their English experience. They all had learned English as an L2 since primary school through classroom instructions, but none of them had

Table 1: Participants’ language background.

Group assignment	Age	Age of L2 acquisition	CET-4 score	Gender
Exp_V1–C1, V2–C2, V3–C3	18.40 (0.44)	8.24 (2.28)	543.39 (48.80)	7 F; 70 M
Exp_V1–C2, V2–C3, V4–C1	18.38 (0.79)	8.98 (3.12)	539.74 (62.19)	12 F; 68 M
Exp_V1–C3, V3–C1, V4–C2	18.31 (0.23)	8.60 (2.71)	540.18 (51.65)	9 F; 71 M
Exp_V2–C1, V3–C2, V4–C3	18.25 (0.43)	8.12 (1.98)	536.45 (26.92)	11 F; 72 M
Con_no_video	18.55 (0.56)	8.32 (2.90)	550.01 (34.24)	10 F; 67 M

Note. C1 = no subtitle, C2 = L2 subtitle, C3 = L1 subtitle. V1, V2, V3, V4 = video versions. Exp = experimental group, Con = control group.

overseas experience or intensive extracurricular English exposure. They had passed the College English Test 4 Band 4 (CET-4), which was the largest national-wide English proficiency test in China with an established reliability and validity (e.g., Jin and Yang 2006). Since the CET-4 test was taken just two weeks prior to the study, it was representative of participants' proficiency at the time of the study. On average, the participants' L2 proficiency measured by CET-4 was 543.33 points out of 710 ($SD = 53.16$), which was equivalent to the Common European Framework of Reference B2 level, based on a recent alignment scheme (Jin et al. 2022).

2.2 Materials

The input stimuli came from four short videos, covering a range of genres and speech styles. Each video lasted about 60 s, which fell in the 30–120 s range for learning (Thompson and Rubin 1996). The first video (1'17") was an excerpt from the Tedtalk "Inside the Mind of a Master Procrastinator" (V1), the second video (1'10") was a scene from the TV episode "Girl Meet Sneak Attack" (V2), the third video (1'08") was a conversation from the cartoon "Arthur" (V3), and the fourth video (1'13") was a narration from the cartoon "Phineas and Ferb" (V4). Contents of the four videos are provided in Appendix B. These clips were chosen to ensure that they could represent as varied and dense connected speech devices as possible within a limited time period. The four videos contained 187, 148, 111, and 199 words, within which the target connected speech occupied 16, 17, 14, and 15 words, respectively. Each video tested six to nine connected speech features.

The target connected speech items are tabulated in Table 2. The items were labeled as linking (LIN), modification (MOD), reduction (RED), contraction (CON), and deletion (DEL), according to the way they were uttered in the videos. The coding was double-checked by another researcher who was also an English teacher with systematic phonological knowledge. One item often adopted multiple connected speech processes. For instance, the phrase "*would like to*" was pronounced as [wə laɪk tə], which involved the deletion of the final consonant [d] in "*would*", as well as the reduction of vowel from [u] to [ə] in "*would*" and "*to*". These items were coded into label combinations (e.g., DEL_RED). In total, 40 % of the tested speech stream used single connective devices, 37 % of them used double devices, and 23 % triple devices. The number of connective devices contained in an item was coded into 1, 2, and 3, which was referred to as the "complexity" of a connected speech item.

To manipulate task conditions, English and Chinese subtitles were added to the videos using the ArcTime subtitle editing software. All the videos were in 1080p resolution and were played to the experimental group on four 70 in. LCD displays on the four walls of a small classroom. Soundtracks (44.1 kHz/24 bps) of the videos were

Table 2: A list of connected speech items and participants’ mean accuracy on each item.

Content	Type	Complexity	Accuracy
I have <i>some</i> good news	RED	1	96 %
Missy invited me to see a movie <i>with her</i>	RED	1	85 %
<i>You are</i> the best daughter	LIN	1	83 %
Come along <i>with us</i>	LIN	1	82 %
So that’s <i>what</i> we did for that day	RED	1	81 %
So we <i>can</i> get a better feel for India	RED	1	69 %
Do you really <i>want your</i> friends to know that our father is their ... Garbage man?	LIN	1	66 %
Come on down! You are missing <i>all the fun!</i>	LIN_RED	2	65 %
Maybe you guys <i>would like to</i>	RED_DEL	2	51 %
Now I certainly appreciate your wanting me to <i>take care of</i> me during a scary movie,	LIN_RED	2	49 %
I think we’d have more fun <i>just</i> hanging out together	RED	1	47 %
The people of Ted <i>decided to</i> release the speakers	LIN	1	35 %
<i>We’ve been</i> spending the last several weeks studying careers	RED_CON	2	33 %
Any mother <i>could</i> dream of having	RED	1	32 %
No <i>need to</i> make her feel worse	LIN	1	29 %
<i>Been</i> a dream of mine to have done a Ted talk in the past	RED	1	26 %
Have volunteered <i>to show us</i> around their places of work	LIN_RED	2	24 %
I said yes. <i>It’s</i> always	RED_CON	2	23 %
And the Monkey said totally agree <i>but also let’s</i> just open google earth	LIN_RED_DEL	3	18 %
Well <i>at least</i> the self-parking works	LIN_RED	2	15 %
These are my friends, and I <i>don’t like</i> doing anything without my friends.	LIN_RED_CON	3	12 %
Have something else <i>in his</i> mind	RED_DEL	2	10 %
<i>Way to go</i> , dad	LIN_RED	2	8 %
But in the middle <i>of all this</i> excitement the rational decision maker seem to	LIN_MOD	2	8 %
All the times I called you delusional and <i>mocked you to</i> my friends behind your back	LIN_RED_MOD	3	7 %
And you showing me your <i>leg and all</i> , but ...	LIN_RED	2	6 %
Well <i>what’s all this</i> yelling about?	LIN_RED_MOD	3	6 %
All those journals I filled <i>with an eye</i> towards a standup comedy	LIN_RED_DEL	3	4 %
We are <i>gonna scroll up</i> for 2.5 h till we get to the top of the country,	LIN_RED_CON	3	1 %
We <i>rotate out</i> with the board of selectmen	LIN_RED_DEL	3	1 %

delivered through four loudspeaker sets placed approximately 1.5 m in front of the participants. Three input conditions of the experiment group were formed: a) English videos without subtitles (C1: no subtitle condition); b) English videos with English subtitles (C2: L2 subtitle condition); c) English videos with Chinese subtitles

(C3: L1 subtitle condition). The control group without watching any video formed the baseline condition (C0: no video condition).

Each video came with a perception test evaluating L2 speakers’ perception of English connected speech. The perception test was a spot dictation task, which visually presented six to nine sentences from the video and left out the target connected speech items such as “*all this*” and “*let’s*” for participants to fill in. The task examined the perception of 30 connected speech items (see Table 2). There were four versions of the spot dictation task, each testing six to eight connected speech items in V1, V2, V3, and V4. Every experimental group would take three of the four versions, depending on the videos they watched.

Four videos, four tests, and three subtitle conditions were created for the experiment group. The materials were divided into four counterbalanced lists (i. V1–C1, V2–C2, V3–C3; ii. V1–C2, V2–C3, V4–C1; iii. V1–C3, V3–C1, V4–C2; iv. V2–C1, V3–C2, V4–C3;) using a Balanced Incomplete Latin Square design (BILS) (Ai et al. 2013). Each list contained a random order of three videos with three subtitle conditions. The four lists ensured that each experimental group saw three different videos in no subtitle condition, L1 subtitle condition, and L2 subtitle condition. There were no filler trials.

2.3 Procedure

The experiment took place in batches of 20–30 students in quiet classrooms. The experimental procedure was schematized in Figure 1. The experimental group of 320 students were first informed that they needed to watch several English videos and there might be English or Chinese subtitles. They were told to pay attention to the

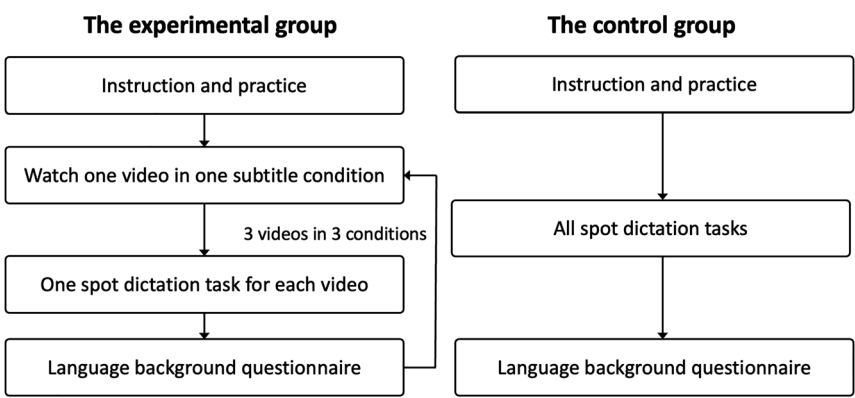


Figure 1: The experimental procedure.

content and the language, as there would be a spot dictation task following each video. Before the experiment, participants had a practice session, during which they watched a short excerpt (32") from "The Simpsons" and did a spot dictation task requiring them to write down the missing connected speech items "*just stupid*" and "*you don't want*" in the excerpt. Participants were allowed to ask clarification questions. The practice session familiarized them with the experimental procedure and made sure that they could hear the sounds clearly.

The experiment group watched three videos, each followed by a spot dictation task to test L2 speech perception. The experiment was administered in the Rain Classroom, a platform that could present multi-media tests and automatically collect written answers (Wang 2017; Wang and Shen 2020). In the perception test, participants were visually provided with the testing sentences, which were transcripts of the lines in the video. They then listened twice to the original recordings of the sentences spoken in the video and spent less than 2 min writing down the missing items in each sentence. After the test, participants completed an online questionnaire about their language background. The experiment lasted 30–40 min.

The control group of 77 students completed all of the spot dictation tests that had been conducted on the experimental group. The testing procedures were the same, except that the control group did not watch any video before the tests. The control group spent 15–20 min to complete the experiment.

2.4 Data analysis

Participants' accuracy was examined by determining whether they could write down the connected speech items in the videos. If their written words were identical to the model answers, their responses were coded as "1", and otherwise "0". This matching was done automatically with Excel functions and then manually checked one by one. The correct answers could be more than one, allowing for possibilities of contraction (e.g., "*we've been*"/"*we have been*") and capitalization (e.g., "*Way to go*"/"*way to go*"). Minor miss-spellings were not coded as errors (e.g., "*we'v been*"), as the purpose was to test speech perception instead of writing, but if a miss-spelling affected meaning (e.g., "*we've been*" → "*we've bin*"), it was coded as incorrect.

Mixed-effects logistic regression models were constructed on the participants' accuracy on the spot dictation tasks (Dixon 2008; Jaeger 2008; Quené and van den Bergh 2008). We chose logit mixed effects models for statistical analyses since accuracy data are binary variables (correct or wrong), which can lead to wrong interpretations when analyzed with the widely used ANOVA (see Jaeger 2008). Mixed logit models are based on binomial distributions and can model binary outcomes like accuracy. Compared with standard logistic regression, mixed logit models can

include both fixed and random effects within one analysis. We fitted the models using the statistical software R (R Core Team 2015) and the lme4 package (Bates et al. 2015). Group comparisons for each condition were calculated using post-hoc tests with the help of the lsmeans package (Lenth 2016).

The model on participants' performances in the spot dictation task included the fixed effects of Condition (C0 = control group, C1 = no subtitle, C2 = L2 subtitle, and C3 = L1 subtitle; baseline = C0), connected speech Complexity (1, 2, and 3; baseline = 1), participants' language Proficiency (as measured in CET scores). Whether a connected speech item commonly occurs in daily life could potentially affect L2 speech perception (Wong et al. 2020), so the model added the fixed effect of Frequency of occurrence. The frequency of each connected speech item was obtained by searching the exact matching word sequence in the Corpus of British English (Davies 2004) and the Corpus of Contemporary American English (COCA) (Davies 2008), including both contracted (e.g., it's) and full forms (e.g., it is).

For interaction terms, the model included Proficiency $\times$ Complexity and Proficiency $\times$ Condition interactions to examine whether high- and low-proficiency L2 speakers behaved the same on connected speech items of various complexities and across different conditions. Complexity $\times$ Condition interaction was entered to show whether the performances on simple and complex connected speech items differed across conditions. Participant and item were entered as crossed random effects, adjusting for both the intercept and the slope of the within-group factor (see Cunnings 2012). The numeric variables, including Frequency, Complexity, and Proficiency were z-score standardized to avoid convergence errors in the estimation of parameters due to the different scales.

Subsequently, the above fixed effects were removed from the initial most complex model once at a time, and model comparisons were run pairwise to determine if the reduced model affected total accounted variances. Model selection was based on Akaike's information criterion (AIC), the Bayesian information criterion (BIC), and the log-likelihood ratio (LogLik). Only the minimal model (Table 3) was reported in the paper. Post-hoc tests with Tukey adjustment were conducted to examine interaction effects.

3 Results

3.1 Facilitative effects of subtitled videos on L2 connected speech perception

The mixed-effects logit model (Table 3) suggested that the perception of connected speech differed in Condition. The accuracies in C1 (videos without subtitles), C2

Table 3: Mixed effects binomial logistic regression analyses on the perception of all the connected speech items.

	β	SE	z Value	p Value
(Intercept)	−1.82	0.25	−7.21	<0.001***
Condition C1	0.46	0.14	3.26	0.001**
Condition C2	0.88	0.14	6.20	<0.001***
Condition C3	0.95	0.14	6.72	<0.001***
Complexity	−1.53	0.22	−6.81	<0.001***
Proficiency	0.19	0.11	1.73	0.084
Condition C1: proficiency	0.10	0.13	0.77	0.442
Condition C2: proficiency	0.15	0.13	1.09	0.274
Condition C3: proficiency	0.37	0.13	2.77	0.006**

(L2 subtitled videos) and C3 (L1 subtitled videos) conditions were all significantly higher than the C0 (no video) condition. Thus, the experimental group, who watched videos before the perception test, were better at perceiving the connected speech than the control group, who did not watch videos before the perception test. The performances of the experimental and the control groups are shown in Figure 2. The effect of Frequency was not significant. Thus, the occurrence of items in BNC and COCA corpora did not significantly influence L2 speakers’ performance.

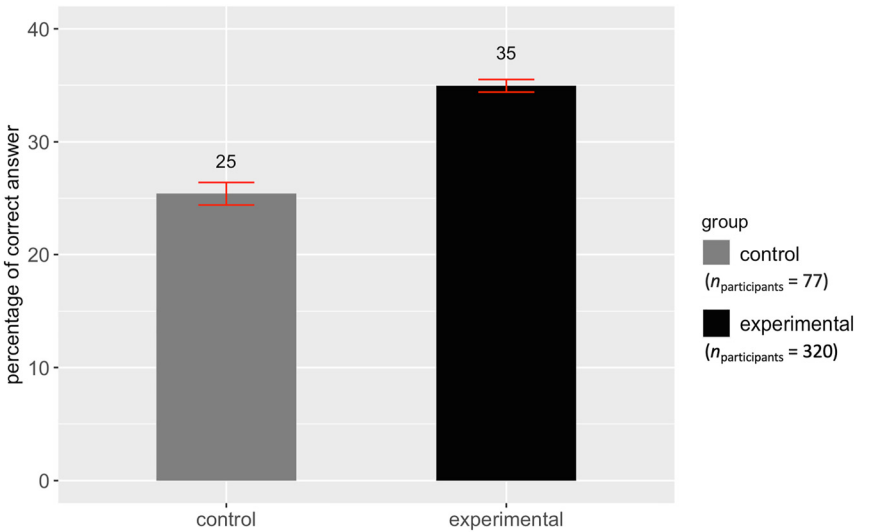


Figure 2: Percentage of correct answers in the spot dictation tasks by participant group. Error bars represent standard error.

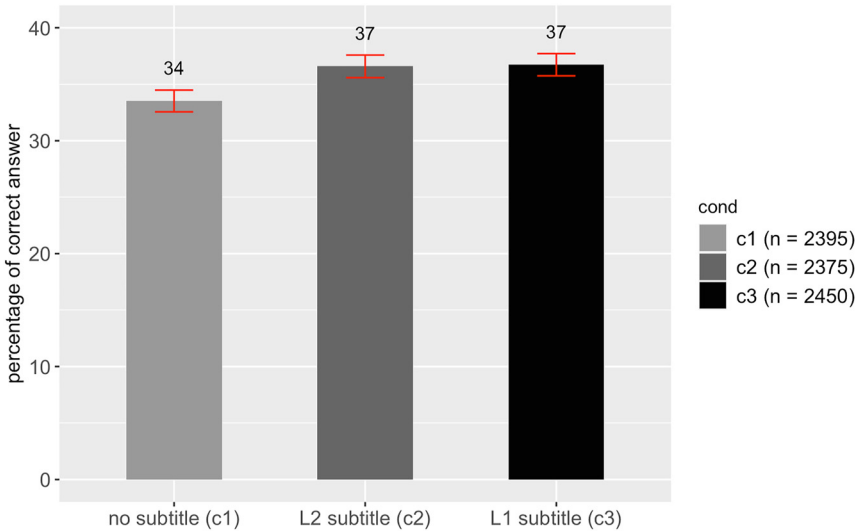


Figure 3: Percentage of correct answers in the spot dictation tasks by subtitle condition. Error bars represent standard error.

Regarding the subtitle effects, post-hoc pairwise comparisons with Tukey adjustments suggested that connected speech perception performances were significantly higher in C2 condition ($n = 2,375$, $mean = -0.95$) than in C1 condition ($n = 2,395$, $mean = -1.36$) ($p < 0.001$), whereas the performances were not significantly different between C2 condition ($n = 2,375$, $mean = -0.95$) and C3 condition ($n = 2,450$, $mean = -0.86$) ($p > 0.05$). The experimental groups' mean accuracies of the perceptual task under no subtitle (C1), L2 subtitle (C2) and L1 subtitle (C3) conditions are summarized in Figure 3. The pattern indicated that both L1 and L2 subtitles benefited L2 video-based speech perception.

3.2 Facilitative effects of language proficiency on video-based connected speech perception

The model in Table 3 also exhibited a significant Condition $\times$ Proficiency interaction. Post-hoc tests indicated that the influence of proficiency on video-based connected speech perception was larger in C2 ($n = 2,375$, $mean = -0.96$) and C3 ($n = 2,450$, $mean = -0.86$) conditions than in C1 ($n = 2,395$, $mean = -1.36$) condition ($p < 0.01$), while the proficiency effect did not differ between C2 and C3 conditions. Figure 4 plots participants' performances in the spot dictation task against their corresponding L2

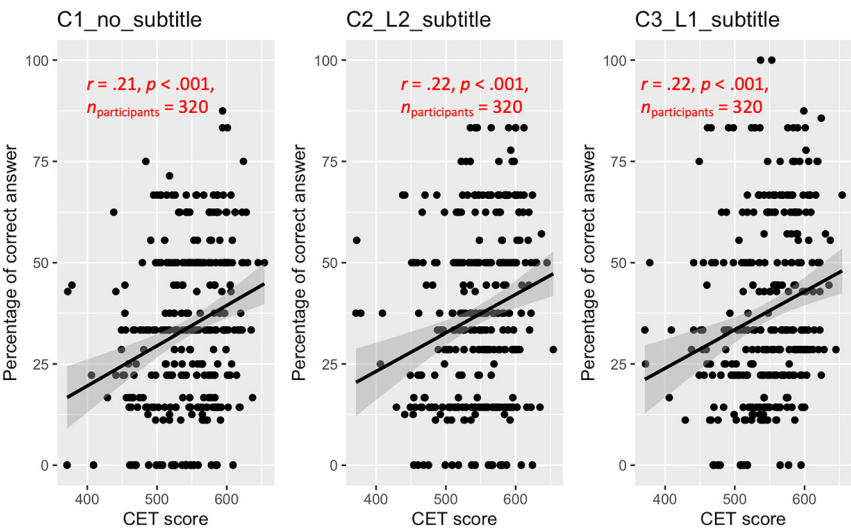


Figure 4: Correlation between L2 speakers’ language proficiencies (measured in CET score) and their performances in the connected speech perception task.

proficiency scores. Pearson’s correlation tests suggested a significant positive correlation between proficiency and connected speech performance in C1 ($n_{\text{participants}} = 320, r = 0.21, p < 0.001$), C2 ($n_{\text{participants}} = 320, r = 0.22, p < 0.001$) and C3 ($n_{\text{participants}} = 320, r = 0.22, p < 0.001$) conditions. Therefore, the more proficient the L2 speakers were, the better they could perceive connected speech in videos. This proficiency effect was slightly more salient when the videos were presented with subtitles.

3.3 The influence of connected speech type on L2 perception

Furthermore, the model (Table 3) revealed a significant main effect of connected speech Complexity on L2 perception. As shown in Figure 5, the more connected speech devices an item used, the less likely it was correctly perceived. The accuracies of each type of connected speech can be found in Table 2.

To investigate the connected speech type effects, three mixed-effects logit models were separately built on accuracies of items using single, double, and triple connected speech devices with Type, Condition, Proficiency, Frequency, and their interactions as fixed effects, and participant and item were entered as crossed random effects. Several of the fixed effects were later stepwise removed as they were insignificant predictors.

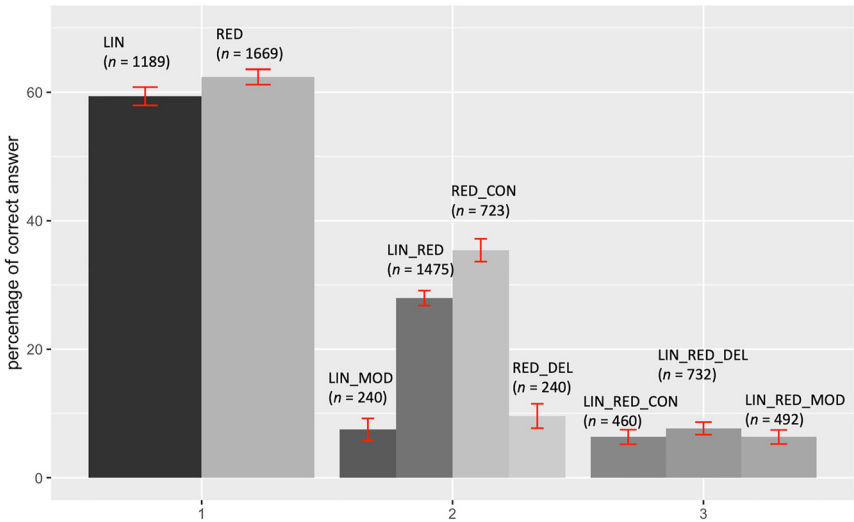


Figure 5: Mean accuracies of all the tested connected speech types grouped with complexity (1 = single connected speech devices, 2 = double connected speech devices, 3 = triple connected speech devices; LIN = linking, RED = reduction, CON = contraction, DEL = deletion, MOD = modification). Error bars represent standard error.

Table 4 displays the output of the model on items with single connected speech devices. Proficiency and Condition were significant predictors of L2 perceptual performance. Higher language proficiency predicted better perception of phrases with single connective devices. As for the Condition effect, post-hoc pairwise comparisons with Tukey adjustment showed that videos with no subtitles (C1, $n = 966$, $mean = 0.35$), L2 subtitles (C2, $n = 920$, $mean = 0.76$), and L1 subtitles (C3, $n = 972$, $mean = 0.78$) all elicited better perception, in comparison to the baseline condition of not watching videos (C0, $n = 756$, $mean = -0.17$). There was no significant Type effect, showing that RED and LIN methods did not lead to different performances.

Table 4: Mixed effects binomial logistic regression analyses on the perception of items with single connective devices.

	β	SE	z Value	p Value
(Intercept)	-0.17	0.40	-0.43	0.671
Condition C1	0.52	0.16	3.25	0.001**
Condition C2	0.93	0.16	5.68	<0.001***
Condition C3	0.96	0.16	5.93	<0.001***
Proficiency	0.38	0.06	6.52	<0.001***

Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 5 shows the results of the model on items with double connected speech devices. The effects of Condition and Proficiency were also significant. Language proficiency positively correlated with the L2 perceptual performance. Yet, the Condition effect was only significant at C2 and C3 levels. Post-hoc pairwise comparisons suggested that L1 (C2, $n = 846$, $mean = -1.19$) and L2 subtitled videos (C3, $n = 880$, $mean = -1.24$) improved L2 speech perception over the baseline condition (C0, $n = 693$, $mean = -1.98$). However, videos with no subtitles (C1, $n = 952$, $mean = -1.78$) did not display this improvement. The Type effect was not significant, and there were no significant performance differences among the items using LIN_MOD, LIN_RED, RED_CON, or RED_DEL methods.

Table 6 shows the output of the model on items with triple connected speech devices, with significant Condition and Proficiency predictors. Speakers with higher proficiency were better at perceiving such connected speech items. Interestingly, only the L1 subtitled video condition (C3, $n = 598$, $mean = -2.94$) yielded higher performance compared with the baseline condition (C0, $n = 441$, $mean = -4.09$), while the videos with L2 subtitles (C2, $n = 609$, $mean = -3.63$) and without subtitles (C1, $n = 477$, $mean = -3.58$) did not show such an advantage. The effect of connected speech Type was not significant, and there were no significant differences in the performances of items with LIN_RED_CON, LIN_RED_DEL, and LIN_RED_MOD methods.

Table 5: Mixed effects binomial logistic regression analyses on the perception of items with double connective devices.

	β	SE	z Value	p Value
(Intercept)	-1.98	0.38	-5.22	<0.001***
Condition C1	0.20	0.17	1.20	0.229
Condition C2	0.79	0.17	4.79	<0.001***
Condition C3	0.74	0.17	4.50	<0.001***
Proficiency	0.34	0.06	5.82	<0.001***

Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 6: Mixed-effects binomial logistic regression analyses on the perception of items with triple connective devices.

	β	SE	z Value	p Value
(Intercept)	-4.09	0.47	-8.69	<0.001***
Condition C1	0.51	0.34	1.48	0.139
Condition C2	0.46	0.34	1.35	0.177
Condition C3	1.15	0.33	3.52	<0.001***
Proficiency	0.26	0.11	2.41	0.016*

Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

The above analyses suggested that the more connected speech devices an item used, the less likely it was correctly perceived by L2 speakers. Importantly, the L2 speakers behaved differently when processing connected speech items of various complexities. For items with single connected speech devices, the L2 speakers could better perceive them by watching videos with and without subtitles. For items with double connected speech devices, the L2 speakers improved their perception by watching videos with L1 and L2 subtitles. For items with triple connected speech devices, the L2 speakers could only enhance their perception by watching videos with L1 subtitles. Thus, the L2 perception was facilitated by subtitles and videos to different degrees depending on the complexity of the connected speech.

3.4 Error patterns

To best capture the difficulties encountered by L2 speakers in perceiving English connected speech, we examined participants' answer sheets and categorized their errors into the following types (see Appendix A error counts).

The first type of error was missing out words in a phrase, which comprised 22 % of the total incorrect responses. For instance, phrases like *“to show us”*, *“want your”*, and *“you to”* were perceived as *“to show/show/show us”*, *“want”*, and *“you”*, respectively. A common feature was that L2 speakers' tended to omit unstressed single-syllabic words (e.g., *“to”* [tə], *“us”* [ʌs], *“you”* [jə]). These words were pronounced in reduced speech, low in loudness, and could easily escape from L2 speakers' attention.

The second commonly occurring error type was misunderstanding (18 %), in which case L2 speakers wrote down a phrase that sounded similar to, but had a completely different meaning from the original testing phrase. These included instances like perceiving *“decided”*, *“with her”*, *“way to go”*, *“all the fun”*, and *“want your”* as *“resided/said/sad/aside”*, *“wither/whether”*, *“we go/where to go/wait to go”*, *“all the phone/on the phone”*, and *“watch your”*, respectively. On the one hand, the phenomenon reflected L2 speakers' imprecise phonetic encoding. For example, the confusion of *“decided”* [dɪsɪd] with *“said”* [sɛd] and *“sad”* [sæd] showed that the speakers could not distinguish between English vowels /æ/, /ɛ/ and /ɪ/. On the other hand, the errors indicated that L2 speakers were unfamiliar with English linking rules, so that *“want your”* [wʌnt jɔr] → [wʌntʃɔr] was misperceived as *“watch your”* [wʌtʃ jɔr].

The third main type of error was the change of wording (12 %). For instance, the connected speech items *“can”*, *“would”*, and *“dream of”* were written down as *“will”*, *“want”* and *“like”*, respectively. It demonstrated a mismatch between L2 speaker' relatively low listening ability and their comparatively high reading ability. In the

experiment, participants saw subtitled videos before the dictation. They comprehended the meanings by reading the lines under the subtitled videos (visual mode), but did not catch the exact speech flow while doing the dictation task (audio mode), so they rephrased the wording while retaining the original meaning in the test.

Other errors included tense/aspect errors (6 %), writing down extra words (1 %), as well as a vast number of irrelevant errors (41 %). Tense/aspect error denoted instances such as the change of “*we’ve been*” into “*we are/we were*”. Extra word error was the addition of unnecessary words, such as the change of “*all the fun*” into “*all of the fun*”. These reflected L2 speakers’ unfamiliarity with English grammar. L2 speakers also wrote down phrases that were completely irrelevant to the testing phrases, and these were categorized as irrelevant errors. For example, they wrote down “I don’t know”, copied words from surrounding sentences, or left the question blank. Irrelevant errors suggested that L2 speakers completely failed to perceive connected speech items during the task. Such large numbers of irrelevant errors showed that our participants had considerable difficulties in perceiving English connected speech.

4 Discussion

The current study explores the extent to which watching subtitled videos could facilitate the L2 perception of English connected speech by conducting video-based spot dictation tasks on 397 L1 Chinese-L2 English speakers.

With regard to RQ 1 about the effects of watching subtitled videos on L2 speech perception, the study overall supports a positive role played by videos. The experiment group, who watched videos before the dictation task, performed better than the control group, who did not watch any video before the task. As for the influence of subtitle form, the study found that both L1 and L2 subtitles improved L2 connected speech perception in videos. The results echo previous studies on L2 listening comprehension (Bird and Williams 2002; Rodgers and Webb 2017; Winke et al. 2010), observing that L2 speakers could improve their overall listening comprehension scores after watching subtitled videos. This study refines the comprehension of general content into connected speech, the latter being low in perceptual prominence and heavily influential on the perception of L2 speech (Field 2008). Past research has provided at least two reasons for the advantage of subtitled videos over non-subtitled videos. First, subtitles help scaffold the auditory input and fill in L2 speakers’ knowledge gap, which improves L2 speech comprehension (Gass et al. 2019). Also, subtitles can attract L2 speakers’ attention to language forms, and when they no longer watch the videos passively, they can actively process and comprehend the speech (Winke et al. 2010). These explanations are applicable

to the present study. When watching subtitled videos, the L2 speakers were aided by the textual information while listening to the connected speech, because they were actively processing the connected speech form due to the presence of subtitles. When watching videos without subtitles, L2 speakers were likely to passively process the content, so they would easily overlook the connected speech because of its low prominence, and there were no texts for them to rely on. This leads to their lower performances in no subtitle condition than in L1 and L2 subtitle conditions.

However, the study did not find conclusive evidence supporting the advantage of L2 subtitles over L1 subtitles, as opposed to Guichon and McLornan (2008) and Hayati and Mohmedi (2011). In this study, L1 and L2 subtitles generated similar accuracy values. What is more, the study found that the L2 speakers could only improve the perception of triple connected speech by watching videos with L1 subtitles. That is, when the connected speech was complex and difficult to perceive, L1 subtitled videos benefited L2 perception more than L2 subtitled videos, rather than the other way around.

It is noticed that the two studies that found L2 subtitles more beneficial than L1 subtitles (Guichon and McLornan 2008; Hayati and Mohmedi 2011) were conducted on French and Persian speakers of L2 English, and they indicated L1 interference to be the driving force for the low performances in L1 subtitle condition. In comparison, L1 interference may not be as strong in this study on Chinese speakers of L2 English. First, there are cultural differences in L2 speakers' subtitle experiences worldwide. Surveys and studies show that French and Iranian audiences in general prefer dubbing while Chinese audiences in general prefer subtitling (Khoshsaligheh et al. 2019; Shevenock 2022; Stoll 2022). Chinese participants in this study might be familiar with L1 subtitled videos and are therefore likely to find L1 subtitles useful instead of distracting. Second, the participants' L1 Chinese is an ideographic language in which ideograms and symbols reflect the ideas, while Persian, French, and English are alphabetic languages in which symbols reflect the pronunciation of the words. Therefore, Chinese texts in L1 subtitled videos are less likely to interfere with L2 English comprehension given the fundamental difference in English and Chinese orthographic-phonetic rules. In fact, many of the participants in our study commented that they felt less anxious when watching L1 Chinese subtitled videos, and anxiety level is reported to be negatively correlated with language achievements (Horwitz et al. 1986). Reduced anxiety could therefore be another cause of the L1 subtitle advantage when the connected speech was complex. Future studies may test possible influences of various emotions on L2 connected speech perception as well.

In terms of RQ 2, the current results show that language proficiency indeed influences L2 connected speech perception, but proficiency does not seem to modulate subtitle effects. Previous findings diverged on the effects of L2 proficiency

on subtitle effects, with some expecting low-proficiency L2 speakers to show more subtitle reliance (Leveridge and Yang 2013; Pujolà 2002), and others assuming L2 speakers of various proficiencies to benefit equally from subtitles (Montero Perez et al. 2013; Winke et al. 2010). The current finding lent support to the latter view. In this study, the performances on connected speech perception increase with speakers' L2 proficiency, but the degree of increase is similar across different subtitle conditions. Speakers regardless of their L2 proficiency could equally benefit from subtitles in video-based speech perception. One reason could be related to L2 speakers' reading ability. Chinese learners of L2 English often demonstrated a stronger reading ability and relatively weaker listening and speaking abilities (Huang 2006). Consequently, it is possible that both high- and low-proficiency L2 speakers in this study possessed enough reading ability to promptly incorporate the incoming subtitles with audio-visual information. Another explanation could be sought from the prevalence of subtitling English films in China (Shevenock 2022; Stoll 2022). Hence, even low-proficiency L2 speakers can watch subtitled English videos without great difficulty because they are accustomed to this way of presentation. Also, it is possible that the proficiency levels in this study are not varied enough. Despite their different CET scores, the participants were all first-year non-English majors with similar L2 learning experiences. Therefore, the current findings could only apply only to this particular group of learners. Further studies on L2 speakers with more diversified linguistic backgrounds may provide a better understanding of the proficiency effects.

As for the connected speech type factor, we propose that it is the connected speech complexity rather than the type that influences L2 speech perception. In our study, nearly half of the connected speech items were completely missed by the L2 speakers, corroborating previous observations that L1 Chinese-L2 English speakers performed poorly in English connected speech production and perception (Liang 2015; Mao and Chen 2013). Contrary to the hypotheses raised in the introduction, the study did not find L1 Chinese-L2 English speakers to perform better on linking than on reduction. The reason could be that there were not enough stimuli with the same complexity that differed with type, and the type effect was overridden by the complexity effect. Studies using controlled stimuli may better solve this question, but the advantage of using natural stimuli is that they can represent the occurrence of connected speech in real life. The study found that natural connected speech often utilized more than one connective device, and the more connective devices a phrase used, the more difficult it is for L2 speakers to understand. This complexity effect can be elucidated by the acoustic difference between the connected speech form and the citation form. For example, the phrase “*need to*” [ni:d tu:] can be pronounced as [ni:tʃu:] if linking and modification are used (double devices), or as [ni:tʃə] if linking, modification and reduction are used (triple devices). Since the

citation form [ni:d tu:] sounded more like to [ni:tfu:] than [ni:tfə], the speech with double devices [ni:tfu:] is easier than the one with triple devices [ni:tfə] to understand. The more connective devices a phrase uses, the less it resembles its citation form in textbooks or dictionaries, and the less likely it is correctly perceived by L2 speakers.

With regard to RQ 3, the study found that the major errors in video-based connected speech included the deletion of words, the misunderstanding of sounds, and the change of wording. This finding is slightly different from Wong et al. (2021b), which is conducted in audio-only mode. In that study, the perceptual errors made by L2 speakers were usually the loss of consonants, vowels, and syllables. Those errors reveal a bottom-up mental process of translating the phonetic signals into meaningful lexical items. In contrast, the current study found that L2 speakers often changed the wording of a connected speech. In those cases, the connected speech was missed but was then recovered by deciphering the overall meaning from the visual, textual, and contextual information on the screen. This implies that a top-down process of filling the lexical gaps is also important in video-based L2 connected speech perception.

The reasons for L2 speakers' failure in connected speech perception are multifaceted. The most direct cause for the erroneous perception is the lack of knowledge in English connective devices, as these structures are not explicitly taught in classrooms and teachers tend to avoid using connected speech in order to make their instructions easily understood. Another important cause for connected speech misperception is the L2 speakers' confusion of English vowels and consonants (e.g., "*all the fun*" [ɔ:l ðə fʌn] is perceived as "*on the phone*" [ʌn ðə fəʊn]), so they need more English phonetic training to increase the sensitivity to these similar sounds. The third reason for the incorrect perception is related to the knowledge of English grammar. Tense and aspect errors are ubiquitous (e.g., "*dream of*" is perceived as "*dreamed of*") in L2 speakers' answers. The unfamiliarity with English grammar also preempts the L2 speakers from guessing the correct phrase from the context. For example, with adequate grammar knowledge, the misperception of "*no need*" into "*no meet*" will never happen, because the sentence "*no need to make her feel worse*" will become the ungrammatical phrase "*no meet to make her feel worse*". To improve L2 speakers' perception of English connected speech, training in English pronunciation, connected speech skills, and grammar are all encouraged. What is more, video-based speech perception involves the process of audiovisual speech integrations, L2 learners should improve their overall language proficiency to free up their cognitive resources that could be used to update misheard information according to context.

5 Conclusion

Through a video-based L2 speech perception experiment, the current study shows that Chinese speakers of L2 English have difficulties in perceiving English connected speech, and that watching subtitled videos is an effective way to assist L2 comprehension. The perception of English connected speech in videos is modulated by L2 proficiency, subtitle form, and the connected speech complexity instead of connected speech type. L2 speakers perform better in watching L1 and L2 subtitled videos than non-subtitled videos, and there is a tendency for L1 subtitles to be more effective than L2 subtitles when the connected speech becomes complex. Furthermore, English connected speech most of the time uses more than one connective device, and the more devices it contains, the more difficult it is for L2 speakers to perceive. Connected speech is crucial for L2 comprehension, as a primary task of understanding a language is to chunk the speech stream meaningfully, so L2 learners, instructors, and researchers are encouraged to allocate more attention to connected speech structures.

Ethics statement: All participants were provided written informed consent in accordance with the Declaration of Helsinki. The ethics committee of the University of Electronic Science and Technology of China approved this study (protocol code #225956), including its participant recruitment procedure and methodology.

Conflict of interest: The authors declare no conflict of interest.

Research funding: This work was funded by Sichuan Federation of Social Science Associations (Award no.: SC23BS019).

Appendix A: Counts of the errors

Item	Missing out words	Wrong words but sound similar	Change wording	Tense or aspect	Extra words	Irrelevant words
It's	98	0	16	9	4	57
Been	0	58	0	65	10	45
All this	129	16	16	0	0	61
In his	48	13	56	0	0	100
But also	87	25	4	0	0	80
let's just						
Gonna	4	7	4	0	0	223
scroll up						
Can	0	0	23	11	1	40
What	0	22	9	0	5	10
Decided	0	53	3	63	3	33

(continued)

Item	Missing out words	Wrong words but sound similar	Change wording	Tense or aspect	Extra words	Irrelevant words
With her	0	21	2	0	0	11
Need to	15	9	4	1	0	128
Would	0	2	38	0	2	65
With us	10	6	4	1	0	18
I don't like	16	36	44	2	4	91
Take care of	44	4	14	0	2	48
Leg and all	72	28	56	0	0	50
Just	0	17	28	0	8	64
Some	0	6	3	0	0	1
We've been	37	10	18	26	0	85
To show us	102	41	37	9	0	12
Way to go	4	167	18	0	0	52
All the fun	12	11	60	33	8	31
Want your	58	13	13	11	0	5
Mocked you to	73	5	29	0	1	122
With an eye	48	36	7	7	0	139
You are	6	7	6	1	7	15
Could	91	10	19	15	2	31
dream of						
All this	52	23	26	6	2	122
Rotate out	1	51	17	1	0	173
At least	3	134	4	0	2	67

Appendix B: Transcription of the testing materials

Video 1

I said, yes, it's [ɪts] always been [bi:n] a dream of mine to have done a Ted talk in the past. But in the middle of all this [əvɔ:lɪs] excitement, the rational decision maker seemed to have something else in his [ɪmɪs] mind. He was saying, are we clear on what we just accepted? Do we get what's going to be now happening one day in the future? We need to sit down and work on this right now. The monkey said totally agree, but also let's [bəʊ:lsəʊlets] just open Google Earth, and let's zoom into the bottom of India, like 200 feet above the ground, and we're gonna scroll up [gənə skrəʊləp] for 2.5 h till we get to the top of the country, so we can [kən] get a better feel for India. So that's what [wɒt] we did that day. And 6 months turned into four and then two and then one. The people of Ted decided to [dɪ saɪdɪdɪtʊ] release the speakers. And I opened up the website and there was my face staring right back at me, and guess who woke up.

Video 2

Speaker A: Missy invited me to see a movie with her [wɪð ər].

Speaker B: I think she's aware of that, Lucas. No need to [ni:tʊ] make her feel worse.

Speaker A: And I was wondering if maybe you guys would like to [wʊ laɪk tə] come along with us [wɪ ðəs].

Speaker C: You want us?

Speaker B: I'm sorry, Lucas. That wasn't the invitation.

Speaker A: Then I'm sorry, Missy. But I can't go.

Speaker B: What are you talking about? Nobody's ever turned me down in my life.

Speaker A: We'll see. These are my friends. And I don't like [aɪ dɒn't laɪk] doing anything without my friends. Right? Riley. I certainly appreciate you want me to take care of [aɪ dɒn't laɪk] you during a scary movie and you showing me your leg and all [lɛgəndʒ ɔ:l]. But back here in the 7th grade, I think maybe we'd have more fun just [dʒəst] hanging out together.

Speaker B: Grow up!

Speaker C: Not yet!

Video 3

Speaker A: I have some [səm] good news. We've been [wɪv bɪn] spending the last several weeks studying careers.

Speaker B: Are we?

Speaker A: So I'm pleased to announce that several of your parents, Murffy's father, Brien's mother, and yes, Francis' father have volunteered to show us [tu ʃəʊəs] around their places of work, which means in short ...

Speaker B: Field trip!

Speaker C: A field trip with dad? Are you out of your mind?

Speaker D: What's wrong with that?

Speaker C: Francine, don't you realize what this means? Have you no pride?

Speaker D: Way to go [weɪtəgəʊ]! Dad

Speaker E: Come on down. You're missing all the fun [ɔ:lðəfʌn].

Speaker C: Do you really want your [wɒntjʊr] friends to know that our father is there? A garbage Man?

Video 4

Speaker A: Oh Candice. All the times I called you delusional and mocked you to [mɒktʃu:tʊ] my friends behind your back. All those journals I filled with an eye [wɪðn aɪ] toward a standup comedy, but you were telling the truth, I'm so sorry. You're [juər] the best daughter any mother could [kəd] dream of having.

Speaker B: Finally, you realize that.

Speaker C: What's all this [watsɔ:l ðɪs] yelling about? My word. That is impressive!

Speaker A: Did you know about this?

Speaker C: No.

Speaker B: What are you gonna do to them? No TV for a month? Or they have to serve me for a year! How about confiscating their toolbox? I know, take away their seats on the city council.

Speaker A: You have seats on the city council?

Speaker E: We rotate out [rə tɜ:təʊ] with the border selectman.

Speaker C: Well some punishment may be in order, but look, there's no harm done, really Ferb.

Speaker E: Here's bolt number 473. I hope that wasn't important. Well, at least [atli:sɪ] the self-parking works.

Speaker A: I'm sorry, but this is really for your own good. We need you both to understand that what you've been doing is way too dangerous. I'd never be able to live with myself if you got hurt. E: We were wearing helmets.

References

- Ai, Mingyao, Kang Li, Senmao Liu & Dennis K. J. Lin. 2013. Balanced incomplete Latin square designs. *Journal of Statistical Planning and Inference* 143(9). 1575–1582.
- Alameen, Ghinwa & John M. Levis. 2015. Connected speech. In Marnie Reed & John M. Levis (eds.), *The handbook of English pronunciation*, 159–174. Hoboken, NJ: Wiley Blackwell.
- Bates, Douglas, Martin Maechler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.
- Bird, Stephen A. & John N. Williams. 2002. The effect of bimodal input on implicit and explicit memory: An investigation into the benefits of within-language subtitling. *Applied Psycholinguistics* 23(4). 509–533.
- Bryła-Cruz, Agnieszka. 2021. The gender factor in the perception of English segments by non-native speakers. *Studies in Second Language Learning and Teaching* 11(1). 103–131.
- Ceuleers, Dorien, Ingeborg Dhooge, Sofie Degeest, Hanneleen Van Steen, Hannah Keppler & Nele Baudonck. 2021. The effects of age, gender and test stimuli on visual speech perception: A preliminary study. *Folia Phoniatrica et Logopaedica* 74(2). 131–140.
- Chen, Ying Yu, Yu Sheng Chang, Jia Ying Lee & Ming Huei Lin. 2021. Effects of a video featuring connected speech instruction on EFL undergraduates in Taiwan. *SAGE Open* 11(2). 1–12.
- Cummings, Ian. 2012. An overview of mixed-effects statistical models for second language researchers. *Second Language Research* 28. 369–382.
- Davies, Mark. 2004. British National Corpus (from Oxford University Press). Available at: <http://www.english-corpora.org/bnc/>.
- Davies, Mark. 2008. The Corpus of Contemporary American English (COCA). Available at: <https://www.english-corpora.org/coca/>.

- Dixon, Peter. 2008. Models of accuracy in repeated-measures designs. *Journal of Memory and Language* 59(4). 447–456.
- Duanmu, San. 2007. *The phonology of standard Chinese*. New York: Oxford University Press.
- Field, John. 2008. Bricks or mortar: Which parts of the input does a second language listener rely on? *TESOL Quarterly* 42(3). 411–432.
- Gass, Susan, Paula Winke, Daniel R. Isbell & Jieun Ahn. 2019. How captions help people learn languages: A working-memory, eye-tracking study. *Language Learning and Technology* 23(2). 84–104.
- Guichon, Nicolas & Sinead McLornan. 2008. The effects of multimodality on L2 learners: Implications for CALL resource design. *System* 36(1). 85–93.
- Hayati, Abdolmajid & Firooz Mohmedi. 2011. The effect of films with and without subtitles on listening comprehension of EFL learners. *British Journal of Educational Technology* 42(1). 181–192.
- Horwitz, Elaine K., Michael B. Horwitz & Joann Cope. 1986. Foreign language classroom anxiety. *The Modern Language Journal* 70(2). 125–132.
- Huang, Jinyan. 2006. English abilities for academic listening: How confident are Chinese students? *College Student Journal* 40(1). 218–226.
- Jaeger, Tim Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59(4). 434–446.
- Jin, Yan, Wei Jie & Wei Wang. 2022. Daxue Yingyu Si, Liuji Kaoshi yu Yuyan Nengli Biaozhun de Duijie Yanjiu [Exploring the alignment between the College English Test and language standards]. *Waiyu Jie [Foreign Language World]* 209(2). 24–32.
- Jin, Jin & Huizhong Yang. 2006. The English proficiency of college and university students in China: As reflected in the CET. *Language, Culture and Curriculum* 19(1). 21–36.
- Khoshsaligheh, Masood, Saeed Ameri & Farzaneh Shokoohmand. 2019. Amateur subtitling in a dubbing country: The reception of Iranian audience. *Observatorio (OBS*) Journal* 13. 71–94.
- Koolstra, Cees M. & Jonannes W. J. Beentjes. 1999. Children's vocabulary acquisition in a foreign language through watching subtitled television programs at home. *Educational Technology Research and Development* 47. 51–60.
- Lenth, Russell V. 2016. Least-squares means: The R package lsmeans. *Journal of Statistical Software* 69(1). 1–33.
- Leveridge, Aubrey Neil & Jie Chi Yang. 2013. Testing learner reliance on caption supports in second language listening comprehension multimedia environments. *ReCALL* 25(2). 199–214.
- Liang, Danxin. 2015. Chinese learners' pronunciation problems and listening difficulties in English connected speech. *Asian Social Science* 11(16). 98–106.
- Lwo, Laurence & Michelle Chia-Tzu Lin. 2012. The effects of captions in teenagers' multimedia L2 learning. *ReCALL* 24(2). 188–208.
- Mao, Huai Zhou & Hua Ying Chen. 2013. Exploring elision of schwa of /ə/ in English utterances by C & U English majors. *International Journal of Applied Linguistics and English Literature* 2(1). 117–125.
- Markham, Paul L., Lizette A. Peter & Teresa J. McCarthy. 2001. The effects of native language vs. target language captions on foreign language students' DVD video comprehension. *Foreign Language Annals* 34(5). 439–445.
- Mayer, Richard E. 2005. *The Cambridge handbook of multimedia learning*. (Ed.) Richard E. Mayer. Cambridge: Cambridge University Press.
- Montero Perez, Maribel, Elke Peters & Piet Desmet. 2014. Is less more? Effectiveness and perceived usefulness of keyword and full captioned video for L2 listening comprehension. *ReCALL* 26(1). 21–43.
- Montero Perez, Maribel, Wim Van Den Noortgate & Piet Desmet. 2013. Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System* 41(3). 720–739.
- Moyer, Alene. 2016. The puzzle of gender effects in phonology. *Journal of Second Language Pronunciation* 2(1). 8–28.

- Paivio, Allan. 1986. *Mental representations: A dual coding approach*. Oxford England: Oxford University Press.
- Paivio, Allan. 2007. *Mind and its evolution: A dual coding theoretical approach*. Mahwah, NJ: Erlbaum.
- Pavlenko, Aneta & Ingrid Piller. 2008. Language education and gender. In Stephen May (ed.), *The encyclopedia of language and education*, 12nd edn. 57–69. New York, NY: Springer.
- Pujolà, Joan Tomàs. 2002. CALLing for help: Researching language learning strategies using help facilities in a web-based multimedia program. *ReCALL* 14(2). 235–262.
- Quené, Hugo & Huub van den Bergh. 2008. Examples of mixed-effects modelling with crossed random effects and with binomial data. *Journal of Memory and Language* 59. 413–425.
- R Core Team. 2015. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Roach, Peter. 2009. *English phonetics and phonology: A practical course*. New York: Cambridge University Press.
- Rodgers, Michael P. H. & Stuart Webb. 2017. The effects of captions on EFL learners' comprehension of English-language television programs. *CALICO Journal* 34(1). 20–38.
- Setter, Jane, Peggy Mok, Ee Low, Donghui Zuo & Ran Ao. 2014. Word juncture characteristics in world Englishes: A research report. *World Englishes* 33(2). 278–291.
- Shevenock, Sarah. 2022. Global streaming: By the numbers. Morning Consult Pro. Available at: <https://pro.morningconsult.com/trend-setters/subtitles-dubbing-streaming>.
- Stoll, Julia. 2022. Audience preferences for subtitles or dubbing 2021, by country. *Statista*. Available at: <https://www.statista.com/statistics/1289864/subtitles-dubbing-audience-preference-by-country/>.
- Sweller, John, Paul Ayres & Slava Kalyuga. 2011. *Cognitive Load Theory*. New York, NY: Springer.
- Teng, Mark Feng. 2022. Incidental L2 vocabulary learning from viewing captioned videos: Effects of learner-related factors. *System* 105. 102736.
- Thompson, Irene & Joan Rubin. 1996. Can strategy instruction improve listening comprehension? *Foreign Language Annals* 29. 331–342.
- Vanderplank, Robert. 2016. *Captioned media in foreign language learning and teaching: Subtitles for the deaf and hard-of-hearing as tools for language learning*. London: Palgrave Macmillan.
- Walker, Robin. 2010. *Teaching the pronunciation of English as a Lingua Franca*. Oxford; New York: Oxford University Press.
- Wang, Shuaiguo. 2017. Rain classroom: The wisdom teaching tool in the context of mobile internet and big data. *Modern Educational Technology* 27(5). 26–32.
- Wang, Xiuli & Shuiyi Shen. 2020. Effects of intelligent teaching tools in colleges and universities: A case study of "Rain Classroom". *Journal of Beijing University of Aeronautics and Astronautics* 33(3). 126–132.
- Winke, Paula, Susan Gass & Tetyana Sydorenko. 2010. The effects of captioning videos used for foreign language listening activities. *Language Learning and Technology* 14(1). 65–86.
- Wong, Simpson W. L., Jessica Dealey, Vina W. H. Leung & Peggy P. K. Mok. 2021a. Production of English connected speech processes: An assessment of Cantonese ESL learners' difficulties obtaining native-like speech. *Language Learning Journal* 49(5). 581–596.
- Wong, Simpson W. L., Vina W. H. Leung, Jenny K. Y. Tsui, Jessica Dealey & Anisa Cheung. 2021b. Chinese ESL learners' perceptual errors of English connected speech: Insights into listening comprehension. *System* 98. 102480.
- Wong, Simpson W. L., Cherry C. Y. Lin, Isabella S. Y. Wong & Anisa Cheung. 2020. The differential effects of subtitles on the comprehension of native English connected speech varying in types and word familiarity. *SAGE Open* 10(2). 1–13.
- Yang, Jie Chi & Peichin Chang. 2014. Captions and reduced forms instruction: The impact on EFL students' listening comprehension. *ReCALL* 26(1). 44–61.